Fixations on Low-Resolution Images

Tilke Judd
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, MA, USA 🏠 ✉

Frédo Durand
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, MA, USA 🏠 ✉

Antonio Torralba
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, MA, USA 🏠 ✉

When an observer looks at an image, his eyes fixate on a few select points. Fixations from different observers are often consistent--observers tend to look at the same locations. We investigate how image resolution affects fixation locations and consistency across humans through an eye tracking experiment. We showed 168 natural images and 25 pink noise images at different resolutions to 64 observers. Each image was shown at eight resolutions (height between 4-512 pixels) and upsampled to 860x1024 pixels for display. The total amount of visual information available ranged from 1/8 to 16 cycles per degree respectively. We measure how well one observer's fixations predict another observer's fixations on the same image at different resolutions using the area under the receiver operating characteristic (ROC) curves as a metric. We found that: 1) Fixations from lower-resolution images can predict fixations on higher-resolution images. 2) Human fixations are biased towards the center for all resolutions and this bias is stronger at lower resolutions. 3) Human fixations become more consistent as resolution increases until around 16-64px (1/2 to 2 cycles per degree) after which consistency remains relatively constant despite the spread of fixations away from the center. 4) Fixation consistency depends on image complexity.

Keywords: eye movement, eye tracking, fixations, low-resolution images, salience, saliency maps, attention, predictability

Introduction

When an observer looks at an image, his or her eyes fixate on a few select points. It is well understood that these fixation locations are heavily influenced by both low-level image features and top-down semantic and task driven factors (Buswell, 1935; Yarbus, 1967; William 1981; Niebur & Koch 1995). However, what is not well known is how the fixations are affected by lowering the resolution of the image.

Some researchers have studied image understanding at low resolution. Harmon & Julez (1973), Bachmann (1991), Schyns & Oliva (1997) and Sinha et al. (2006) have done face perception studies which show that when an image of a face is downsampled to a resolution of 16x16 pixels, viewers are still able to identify gender and emotion reliably. Others have shown that we can essentially understand images, or at least the gist of the images (Friedman, 1979; Wolfe, 1998; Oliva, 2005), at a very low resolution (Potter & Levy, 1969; Potter, 1975; Oliva & Schyns, 2000; Oliva & Torralba 2001; Castelhano & Henderson, 2008). Torralba (2009) showed that viewers can classify the scene of an image and identify several objects in an image robustly even when the image has a spatial resolution as low as 32x32 pixels.

If we understand the gist of the scene at a low resolution, our fixations on low-resolution images are likely directed to locations that we expect to see objects of interest (Loftus & Machworth, 1978; Biederman et al., 1982; De Graef et al., 1990; Henderson et al., 1999). Are these fixations likely to land at the same locations as the actual objects of interest in the high-resolution images? We hypothesize that fixation locations should be similar across resolutions, and more interestingly, that fixations on low-resolution images would be similar to and predictive of fixations on high-resolution images.

As further motivation for our work, we noticed that many computational models which aim to predict where people look used features at multiple scales. Some models that predict where people look are biologically-inspired bottom-up computational models based on multiscale low-level image features (Koch & Ullman, 1985; Itti et al., 1998; Rosenholtz, 1999; Itti & Koch, 2000; Privitera & Stark, 2000; Parkhust et al., 2002; Li, 2002; Parkhurst et al., 2002; Torralba, 2003; Parkhurst & Niebur, 2003; van Zoest et al., 2004; Peters, 2005; Hou & Zhang, 2007). Other models include top down features such as face and person detection (Hershler & Hochstein, 2005 & 2006; VanRullen R., 2006; Cerf et al., 2008), horizon line or context detection (Torralba et al., 2006; Ehinger et al., 2009), text detection, object detection or a combination of many of them (Oliva, 2003; Judd et al., 2009). Others use mathematical approaches to predict fixations (Bruce & Tsotsos, 2006, 2009; Kienzle et al., 2007; Avraham & Lindenbaum, 2009) or natural image statistics (Zhang et al., 2008). Some models also attempt to predict fixations during visual search task where low-level image features have little or no impact on fixations (Tsotsos et al., 1995; Rao, Zelinsky, Hayhoe, & Ballard, 2002; Henderson et al., 2006; Underwood et al., 2006; Navalpakkam & Itti, 2007; Einhäuser et al., 2008). For the design of future models, it is interesting to get a notion as to whether all levels of image features are equally important.

In addition, these computational models are designed to predict where people look in relatively high-resolution images (above 256 pixels per side) and often are created from and evaluated with fixations on high-resolution images (such as the fixation databases of Ramanathan et al., 2010 or Judd et al., 2009). Where people look on low-resolution images is rarely studied, and more generally, how fixation locations are influenced by the resolution of the image is not well understood. In this work, we explicitly study how the resolution of an image influences the locations where people fixate.

In this work we track observers' eye movements in a free-viewing memory task on images at 8 different resolutions. This allows us to analyze how well fixations of observers on images of different resolutions correlate to each other and sheds light on the way attention is allocated when different amounts of information are available.

Methods

Images

The images we used for this study were drawn from the image dataset of Judd et al. (2009). We collected 168 natural images cropped to the size of 860x1024 pixels. As a control, we also created 25 fractal (pink) noise images, with a power spectral density of the form 1/f^(5-2*fractal_dim) where our fractal_dim was set to 1. We chose fractal_dim=1 because it most closely resembles the frequency of natural images (Kayser et al., 2006). For each of the natural and noisy images, we generated eight low-resolution images with 4, 8, 16, 32, 64, 128, 256 and 512 pixels along the height (See Figure 1). To reduce the resolution of each image, we used the same method as Torralba (2009): we applied a low-pass binomial filter to each color channel (with kernel [1 4 6 4 1]), and then downsampled the filtered image by a factor of 2. Each pixel was quantized to 8 bits for each color channel. By low-pass filtering the images, we found that the range of

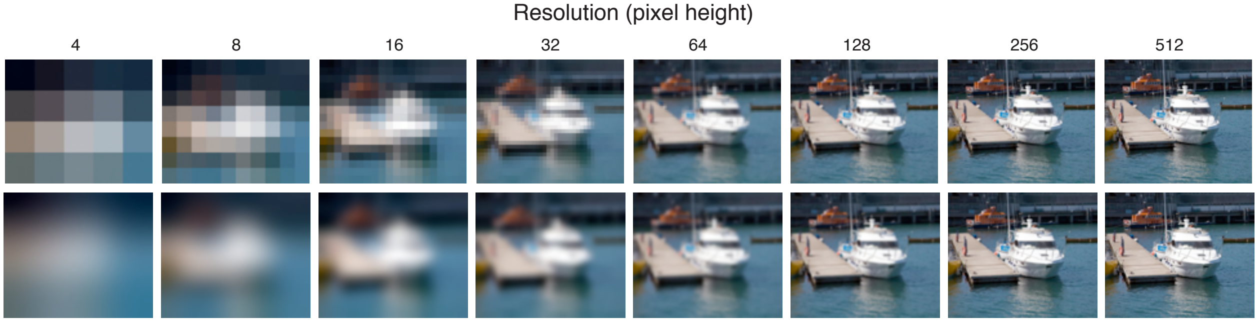


Figure 1 Creating the low resolution images. The two rows of images illustrate the amount of information available at each resolution. The top row shows the downsampled images at each resolution (from 4xN to 512xN), and the second row shows the images upsampled to the original size 860x1024. The upsampled images were shown to participants.

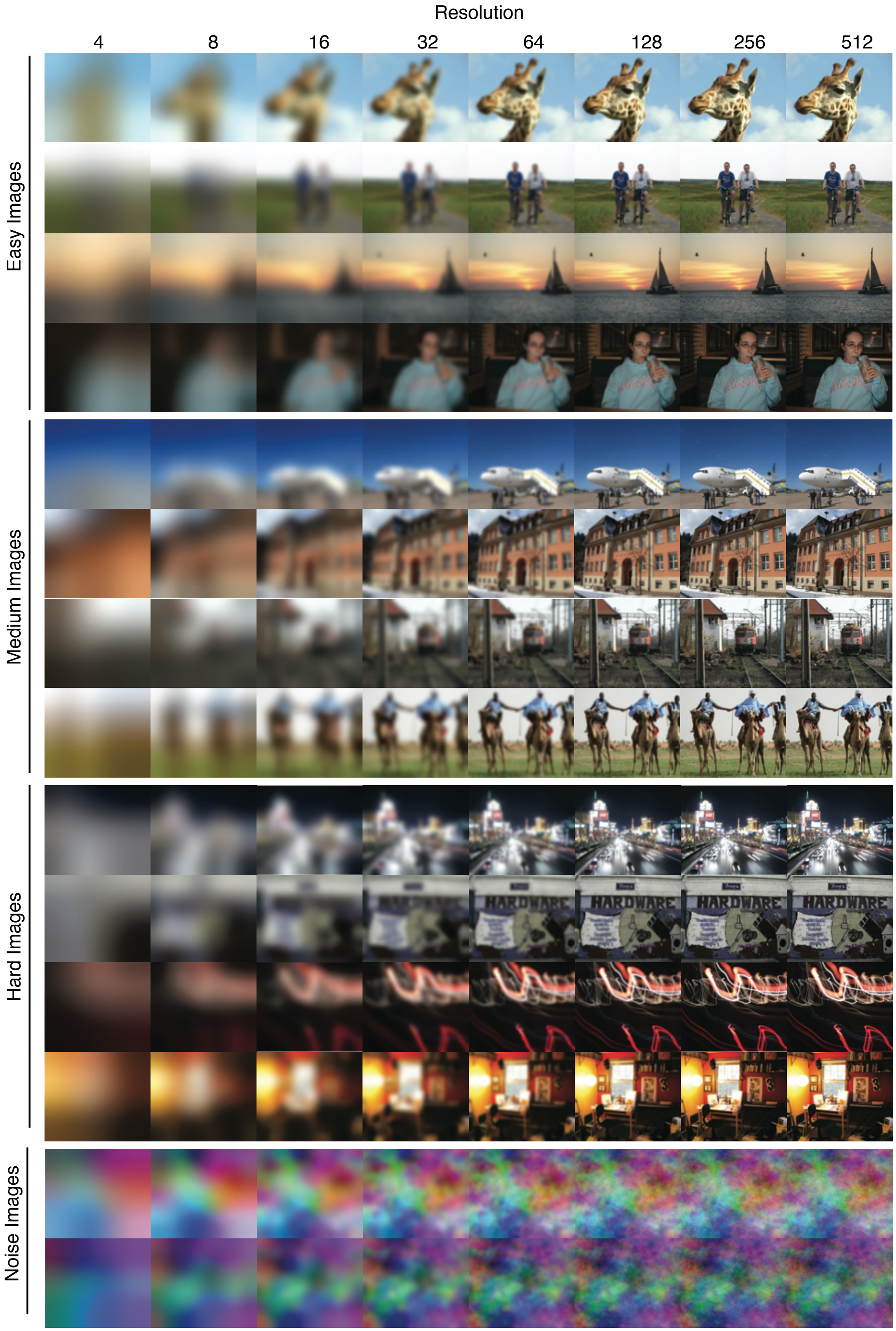


Figure 2 Examples of easy, medium, hard and noisy images used in the eye tracking experiment. Note how more resolution is needed to understand the hard images as compared to the easy images. In addition, hard images at high resolution offer more things to look at.

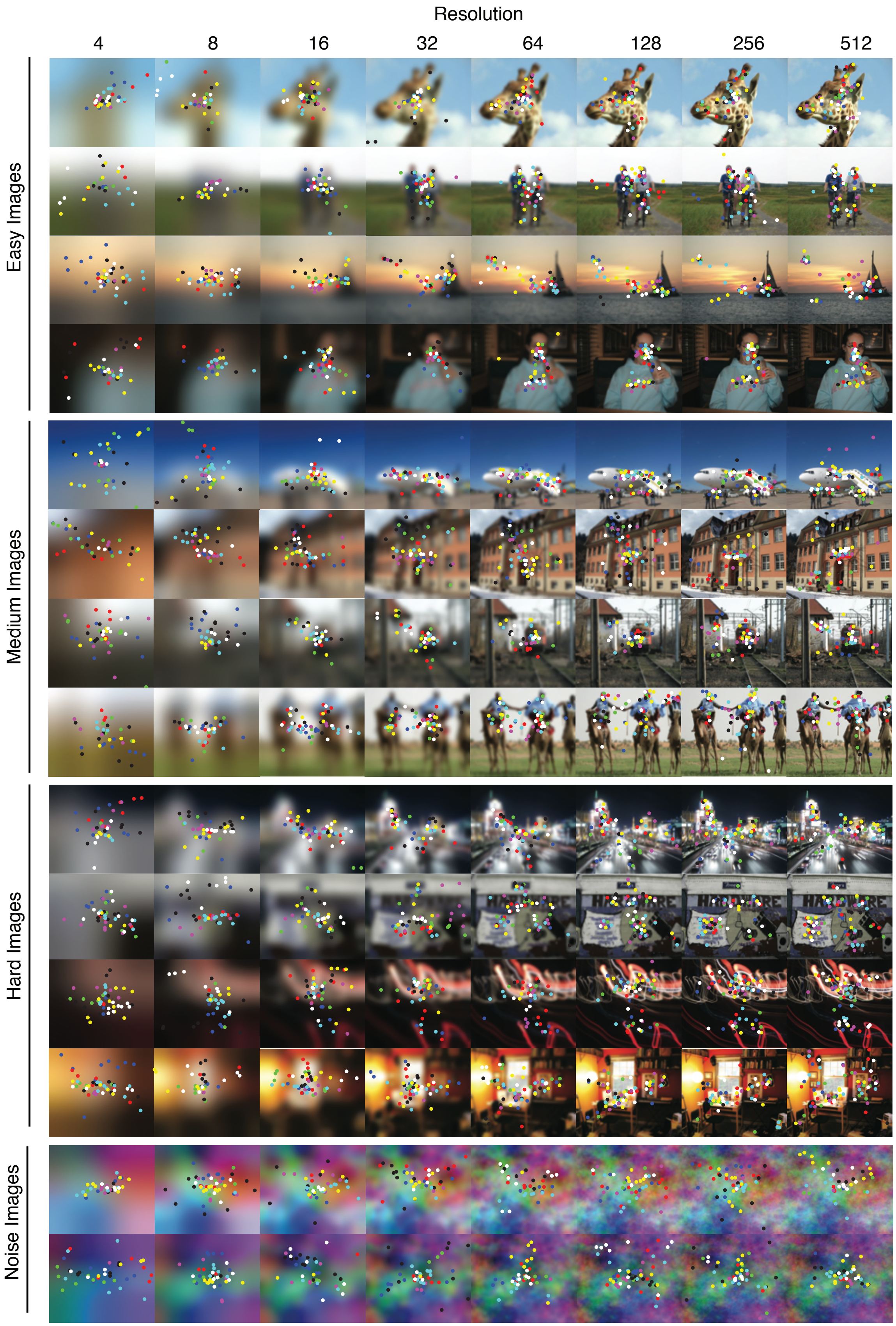


Figure 3 Examples of easy, medium, hard and noisy images used in the eye tracking experiment with all fixations from each of the 8 viewers who saw each image. Note that for the easy images, fixations often remain consistently on the primary object of interest even as the resolution increases. On the other hand, fixations spread out to the many details available on the hard images as resolution increases.

colors was reduced and regressed towards the mean. Since color is an important image feature, we wanted to maintain the range of colors across the blurred versions of a particular image. To do this, we scaled the range of each downsampled image as large as possible within the 0-1 range while maintaining the same mean luminance pixel value. For visualization, the low-resolution images were upsampled using the binomial filter to the original image size of 860x1024 pixels. We used code from Simoncelli's Steerable Pyramid Toolboox to downsample and upsample the images. In total we had 1544 images (193 images at 8 resolutions). In this paper, we use the size of the downsampled image as a measure of the amount of visual information that is available in the blurred images.

In addition we separated the natural images into into easy/medium/hard bins based on their complexity using the following informal criterion: each image was displayed at several resolutions and the author estimated the lowest resolution at which the image could be understood. The images were ranked such that images understood at low resolution were ranked first, and images understood at higher resolutions were ranked last. The ranked list was then binned into three groups of easy, medium and hard images. Easy images tended to contain one large object or simple landscape and could be understood at 16-32px of resolution. Medium images had multiple objects or more complexity and were understood around 32-64px of resolution. Hard images had lots of small details or were often abstract and required 64-128px of resolution to understand. Figure 2 shows a sample of the natural images in the easy, medium and hard categories, and some noise images, all of which we used in our experiment.

Participants

64 observers (35 males, 29 females, age range 18-55) participated in our eye tracking study. Each reported normal or corrected-to-normal vision. They all signed a consent form and were paid $15 for their time. Each observer saw a 193 image subset of the 1544 images and never saw the same image at different resolutions. We distributed the images such that exactly 8 observers viewed each of the 1544 images.

Procedure

All the viewers sat approximately 24 inches from a 19-inch computer screen of resolution 1280x1024px in a dark room and used a chin rest to stabilize their head. A table-mounted, video-based ETL 400 ISCAN eye tracker recorded their gaze path at 240Hz as they viewed each image for 3 seconds. We used a five point calibration system, during which the coordinates of the pupil and corneal reflection were recorded for positions in the center and each corner of the screen. We checked camera calibration every 50 images and recalibrated if necessary. The average calibration error was less than one degree of visual angle (~35pixels). During the experiment, position data was transmitted from the eye tracking computer to the presentation computer so as to ensure that the observer fixated on a cross in the center of a gray screen for 500ms prior to the presentation of the next image. We provided a memory test at the end of the viewing session to motivate observers to pay attention: we showed them 12 images and asked them if they had seen them before. This was not used in the data analysis.

The raw data from the eye tracker consisted of time and position values for each data sample. We use the method from Torralba et al. (2006) to define saccades by a combination of velocity and distance criteria. Eye movements smaller than the predetermined criteria were considered drift within a fixation. Individual fixation durations were computed as elapsed time between saccades and the position of each fixation was computed from the average position of each data point within the fixation. The code for identifying saccades and fixations is on our website at http://people.csail.mit.edu/tjudd/LowRes/Code/checkFixations.m.

We discarded the first fixation from each scanpath to avoid the trivial information from the initial fixation in the center. Figure 3 shows the fixation locations for eight different observers on some of the images used.

Results

We have created an interactive webpage (http://people.csail.mit.edu/tjudd/LowRes/seeFixations.html) which allows readers to view the fixation data collected from our experiment and get an intuitive understanding for where people look on images of different resolutions.

Figure 4 shows that, as the resolution of the image decreases, observers make significantly fewer fixations. Within 3 seconds of viewing time, natural images at a resolution of 512px have an average of 7.9 fixations, while images with 4 pixels of resolution have an average just above 5 fixations [paired t-test:t(167) = 21.2 p<0.001]. We found that 97% of our natural images have an average of at least 4 fixations. Similar trends hold true for noise images. Having more fixations at high resolutions is understandable since high-resolution images have a lot more details that attract the attention of viewers; there is more to look at. On lower resolution images, people seem to dwell in a location either to try and focus, or because they have

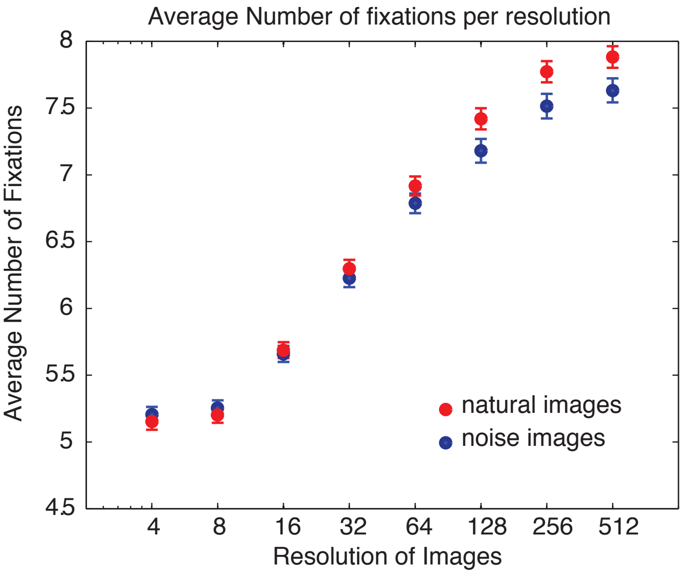


Figure 4 The average number of fixations per viewer in 3 seconds of viewing decreases with lower image resolution. The error bars show the standard error over images.

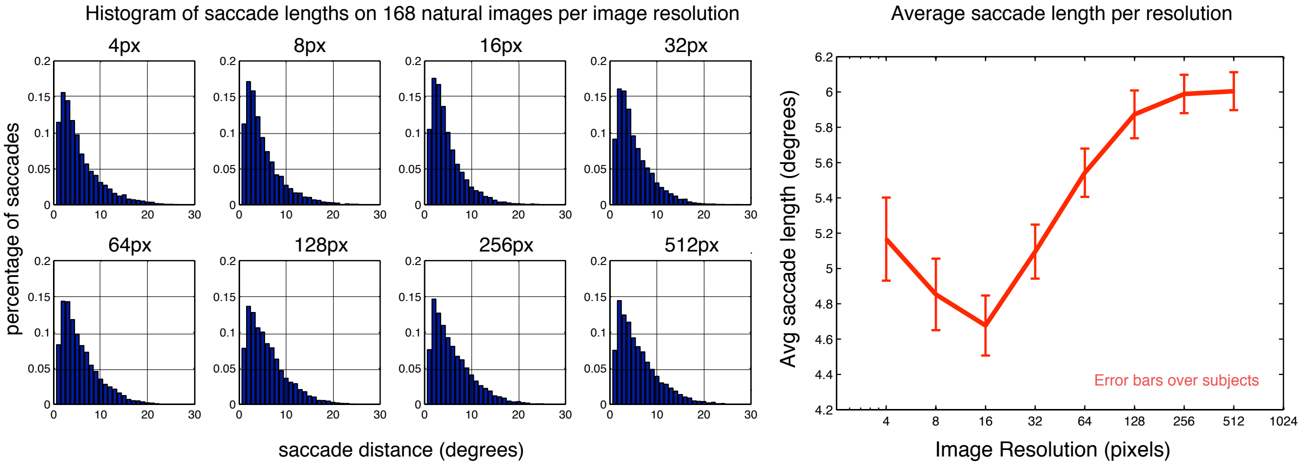


Figure 5 Histogram (left) and plot (right) of saccade lengths across resolutions. For every resolution there are more short saccades than long saccades. The *average* length of saccades increases with resolution.

nothing else to do--nothing in the image is salient enough to pull their attention away from its current location. We did not observe any obvious effect of the memory test on dwell time.

Figure 5 shows a histogram of saccade lengths per resolution on natural images. There is a trend across all resolutions to have many short saccades and fewer long saccades. On average, there are slightly more long saccades on higher resolution images. The plot on the right of Figure 5 shows the average saccade length per resolution, with error bars as standard error over subjects. The mean saccade length increases significantly from 4.6 degrees on 16px resolution to 6 degrees on 128, 256, and 512px resolutions (paired t-test t(63)=10.65 p<0.001). Interestingly there is also a small decrease in saccade length as the resolution increases from 5.2 degrees at 4px to 4.6 degrees at 16px (paired t-test t(63)=3.63 p<0.001).

For each image, we create a fixation map similar to the continuous "landscape map" of Velichkovsky et al (1996) by convolving a Gaussian over the fixation locations of each observer who viewed that image (see the Fixation Map of Image 1 in Figure 7). We choose the size of the Gaussian to have a cutoff frequency of 8 cycles per image, or about 1 degree of visual angle (Einhauser et al., 2008) to match with the area that an observer sees at high focus around the point of fixation. We also made fixation maps with Gaussians of other sizes but found that they did not significantly change the measures of consistency of fixations which we use.

Figure 6 shows the average fixation map of all 168 natural and 25 noise images for each resolution. To measure the quantitative difference between the spread of the fixations across the different fixation maps, we measure the entropy of each fixation map intensity image and add that to each fixation map in Figure 6. Entropy is a statistical measure of randomness used to characterize the texture of the input image. It is defined as -sum(p.*log2(p)) where p contains the histogram fractions for the image intensity divided into 256 bins. The higher the entropy, the more spread out the fixations are. In general, researchers have shown that fixations on high resolution natural images tend to be biased towards the center of the image (Tatler, 2007; Tatler & Vincent 2009; Tseng et al. 2009). Here in Figure 6 we see that as resolution decreases, fixations on natural images get continuously more concentrated at the center of the image. The trend exists for easy, medium and hard natural image subsets. With noise images, fixations remain consistently biased towards the center of the image for all resolutions.

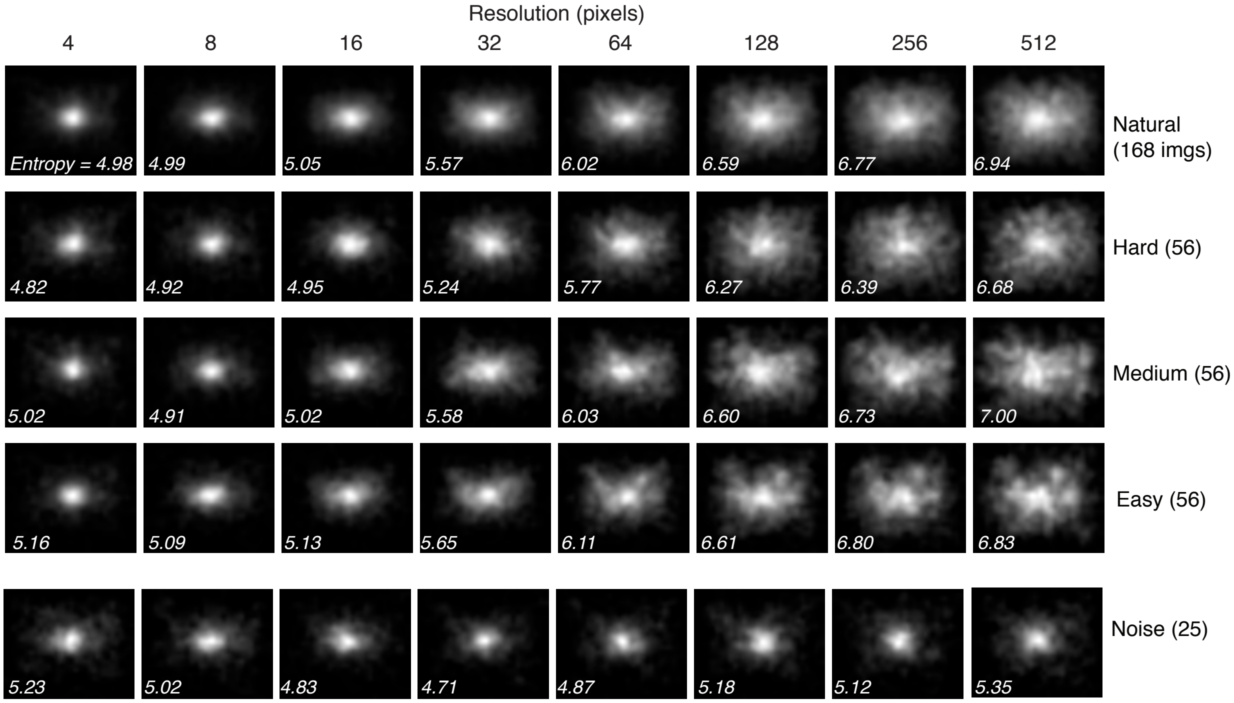


Figure 6 Average Fixation Maps The first row shows the average fixation maps for all 168 natural images for each resolution. In general, as the resolution decreases the fixations become more concentrated at the center. The next three rows show the trend for the easy, medium, and hard subsets of the natural images. The overall trends are the same for each subset. Lastly, the fixation maps for the noise images indicate that fixations are equally biased towards the center independent of the resolution. The entropy of the intensity image for each fixation map is shown in the lower left corner.

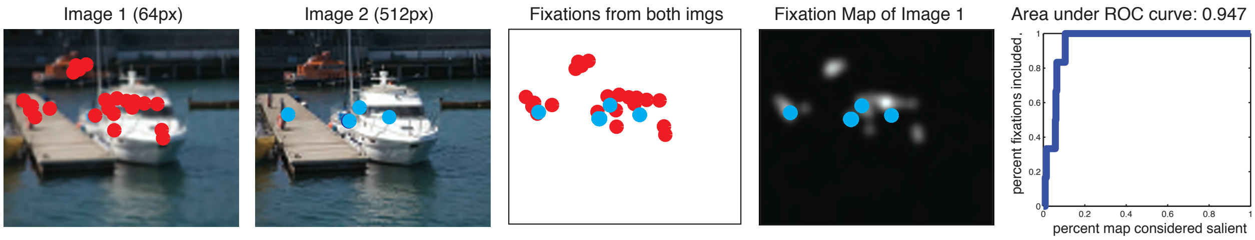


Figure 7 Calculating prediction performance. We use an ROC curve to measure how well a fixation map for an image created from the fixations of several users predict the fixations of a different user on the same image at either the same or a different resolution.

Measuring consistency of fixations

How much variability is there between observers who look at an image at a given resolution, or between observers who look at different resolutions of the same image? To figure this out, we first computed the consistency, or agreement among fixations by the 8 separate observers on the same image of a given resolution (Mannan, Ruddock & Wooding, 1995; Tatler, Baddeley & Gilchrist, 2005). Following the method from Torralba et al. (2006), we measured the inter-observer agreement for each image by using the fixations generated by all-except-one observers to create an "observer-defined" fixation map which was then used to predict fixations of the excluded observer. We use the Receiver Operating Characteristic (ROC) metric to evaluate how well a fixation map predicts fixations from the excluded observer (see Figure 7). With this method, the fixation map is treated as a binary classifier on every pixel in the image. The map is thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated. By varying the threshold, the ROC curve is drawn: the horizontal axis is the proportion of the image area not actually fixated selected by the fixation map (false alarm rate), and the vertical axis is the proportion of fixations that fall within the fixation-defined map (detection rate). In this paper, we report the area under the curve (AUC). This one number corresponds to the probability that the fixation map will rank an actual fixation location more highly than a non-fixated location, with a value ranging from 0.5 (chance performance) to 1 (perfect performance) (Tatler et al., 2005; Harel et al., 2006; Renninger et al., 2007). The code we use for calculating the AUC is available at http://people.csail.mit.edu/tjudd/LowRes/Code/predictFixations.m. For each image at a given resolution, this process was iterated for all observers. The final measure of consistency among the observers for a particular image at a given resolution was an average of 8 AUC values.

Similarly, we also measure the consistency of observers on *different resolutions* of the same image. We use the fixations from all-except-one observers on the image at a first resolution to predict the fixations of one of the observers on the image at the second resolution. This is iterated for all 8 sets of 7 observers of the first image predicting each of the 8 observers of the second image, yielding a final measure which is an average of 64 AUC values.

Not all of the agreement between observers is driven by the image--human fixations exhibit regularities that distinguish them from randomly selected image locations. The fixations from all the images in our database are biased towards the center (see Figure 6). We can measure how centered the fixations are by using a fixation map of a Gaussian of one cycle per image centered on the image to predict observers' fixations. In this "center map" or "center model", the value of a pixel in the map is relative to the distance of the pixel to the center of the map; pixels at the center are highest and pixels on the edges lowest. We can compare the measure of consistency of different observers' fixations with the performance of the center map to predict fixations.

Using the above methods, we now have a way of computing the consistency of fixations among observers on an image at a given resolution, the consistency of fixations across different resolutions of the image, and the performance of the center map to predict fixations. Since we want to know in general how consistent observers are on each resolution, and how consistent fixations are across resolutions, we create what we call a *prediction matrix* per image. The rows and columns of the prediction matrix correspond to the varying image resolution from 4px to 512px. Each entry in the matrix is the average AUC value indicating how well fixations of the image at a given resolution along the row predict fixations of the image at a given resolution along the column, i.e how consistent the fixations of the observers are. The diagonal entries show how consistent observers' fixations are on *an image at a given resolution* (an average of 8 AUC values). The off-diagonal terms measure how consistent observers' fixations are *an image across different resolutions* (an average of 64 AUC values). As a baseline, we also include the performance of how well the center model and the chance model predict fixations on each resolution. The chance model gives a random value

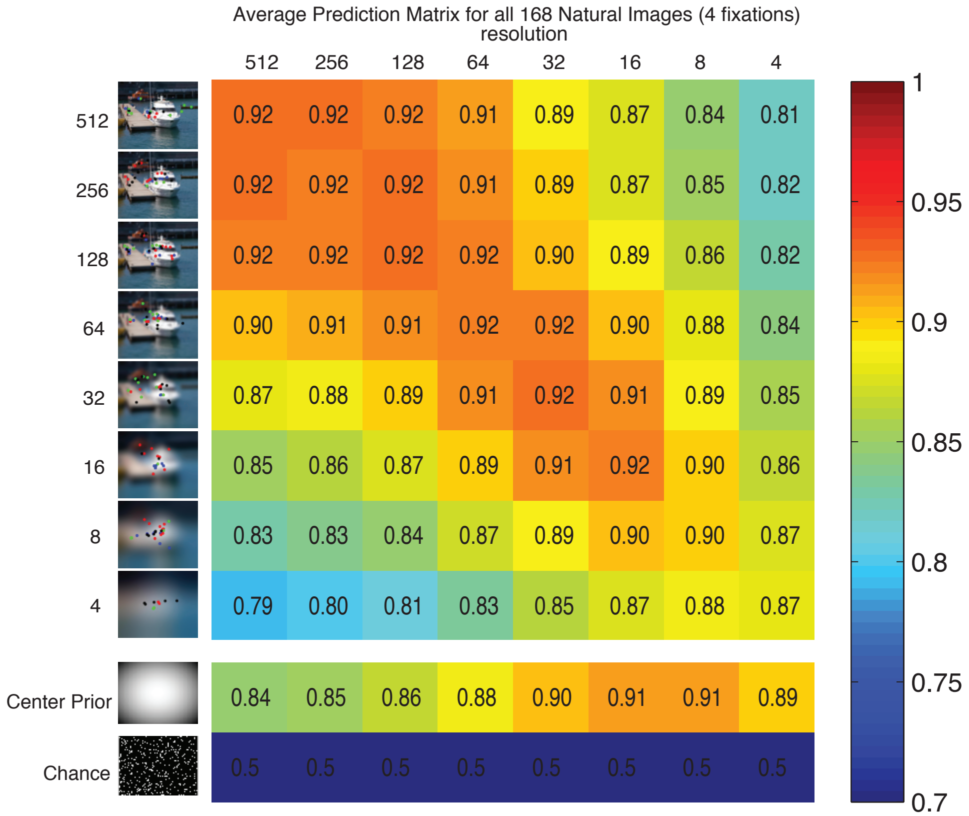


Figure 8 Average Prediction Matrices all natural images. This shows how well fixations of images down the rows predict the fixations of the images along the columns. Diagonal entries show how consistent the fixations of the eight observers are and the off-diagonals show how consistent fixations are between observers on different resolutions of the image.

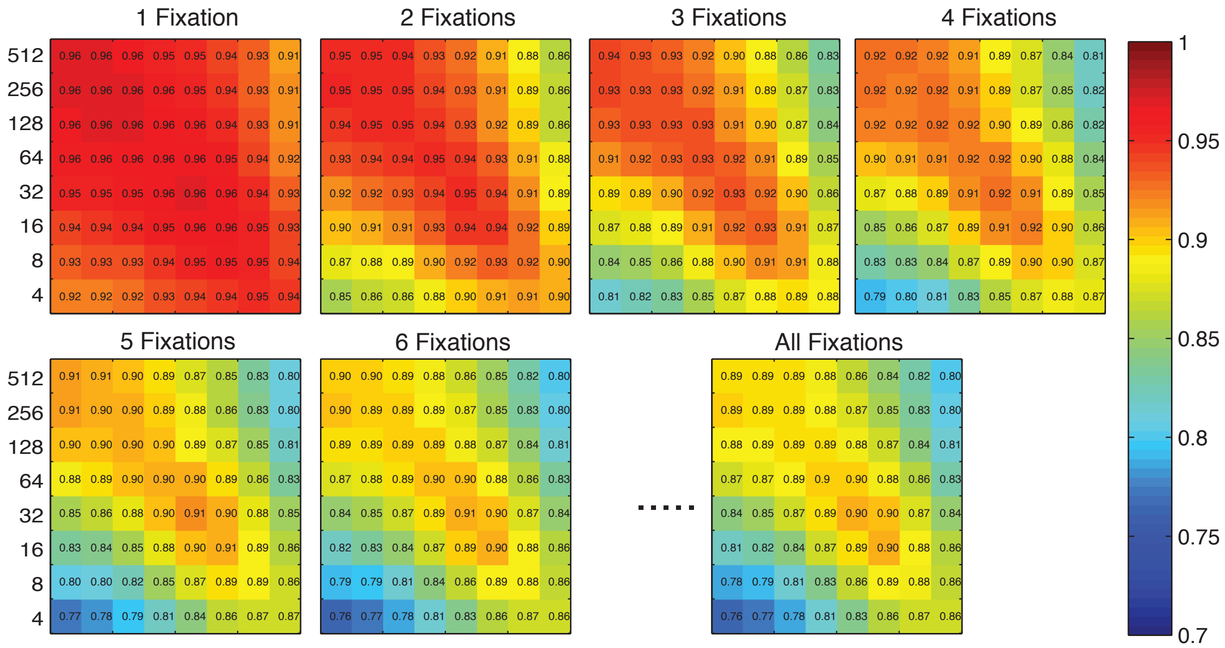


Figure 9 Average Prediction Matrices for different numbers of fixations. The first prediction matrix represents how well the first fixations from a given image resolution predict the first fixations on other resolutions. Note that earlier fixations are more consistent than later fixations. Notice also that fixations on the 32 resolution image are most consistent when many fixations are considered.

between zero and one for each pixel to create a randomly speckled fixation map.

By averaging together the prediction matrix of all 168 natural images in our database, we get the *average prediction matrix* in Figure 8. The average prediction matrix is computed considering the first 4 fixations of each observer. As an example in reading the prediction matrix, note that on average, fixations on high-resolution 512px images predict fixations on 512px images with an AUC=0.92, while fixations on low resolution 4px images predict fixations on the 512px images noticeably less well with an AUC=0.79.

From this average prediction matrix in Figure 8, it is evident that fixations for an image at a specific resolution are best predicted by fixations on the image at the same specific resolution (as seen by the highest average AUC entries along the diagonal). However, it is also evident that fixations on an image can be very well predicted by fixations on images of different resolutions, *including lower resolutions.* In addition, humans fixations are far more consistent than chance, and humans fixations are better at predicting fixations than the center model for all resolutions except the very lowest resolutions (4 and 8px).

Because we would like to see how consistency of earlier fixations are different from consistency among all fixations, we show the average prediction maps for a specific number of fixations (1, 2, 3, 4, 6, 8, all fixations) as in Figure 9.

Discussion

When evaluating our data, we start by asking the following two specific questions:

1) How well do fixations on different resolutions predict fixations on high-resolution images? This corresponds to the first column of the prediction matrices of Figures 8 and 9. We find that fixations on low resolution images can predict fixations on high-resolution images quite well down to a resolution of about 64px. After that performance drops more more substantially but does not drop below the baseline performance of the center map until 16px.

2) How consistent are the fixations across observers on a given resolution? This corresponds to the diagonal of the prediction matrices. We find that consistency varies across resolution. As resolution increases from 4px to 32px, consistency of fixations between humans increases. After around 32px, fixation consistency stays relatively constant despite the spread of fixations away from the center.

In addition, we observe the following trends:

3) Fixations across observers and images are biased towards the center of the image and the bias is stronger as the resolution decreases.

4) Human consistency, or the performance of human fixations to predict new human fixations, is almost always higher than several baseline models and artificial models of saliency showing that humans are the best predictors of other's fixations.

5) Image complexity affects fixation consistency: the more complex the image, the less consistent the fixations. Image consistency on noise images is very poor.

We explore these results more thoroughly in the following discussion.

Fixations on low-resolution images can predict fixations on high-resolution images

Figure 10(a) shows how well fixations from images of each resolution predict fixations on the highest 512px resolution images. The multiple lines represent the AUC values when considering different numbers of fixations per viewer (either 1, 2, 4 or all fixations). Four one-way repeated measure ANOVAs reveal a main effect of the resolution [F(7, 441)={13.8, 35.3, 29.8, 15.4} for {1, 2, 4, all} fixations with every p<0.001]. This indicates that prediction performance increases with resolution independent of the number of fixations considered. The graph also shows that the first fixations are easier to predict than all the fixations in absolute terms.

We look more closely at how prediction performance increases with resolution. Note for example, that when considering 4 fixations per observer, fixations on high-resolution 512px images predict fixations on 512px images (AUC=0.92) significantly better than fixations on low resolution 4px images predict fixations on the 512px images (AUC=0.79) [paired t test: t(63)=19.4, p<0.001], or fixations on 32px images predict

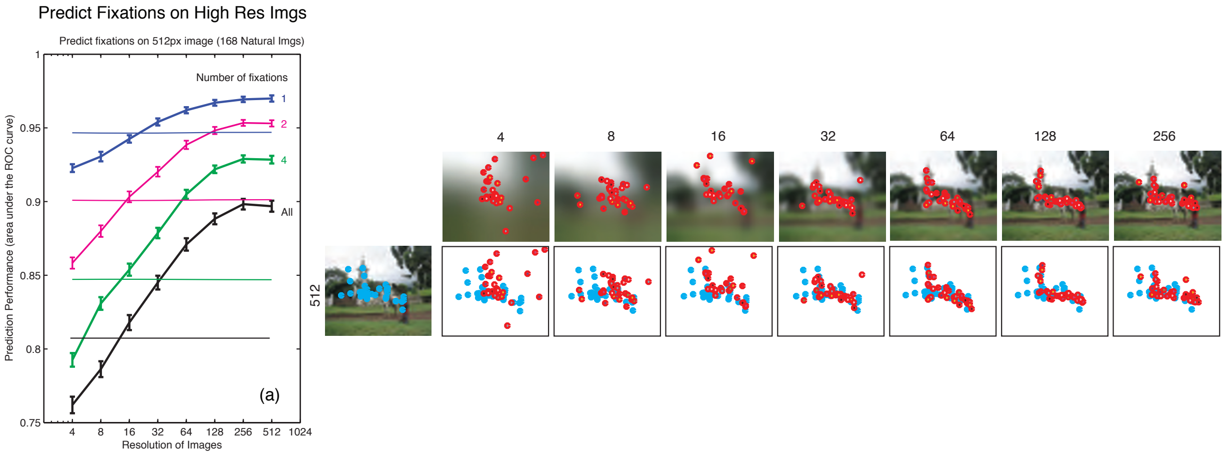


Figure 10(a) Performance on natural images. The graph on the left shows how well the fixations of each resolution predict the fixations on the high resolution images. This is compared to how well the center map (thin line) predicts high resolution images. In general, fixations on images of 16px and above outperform the center map. Performance increases with resolution, and the rate of improvement slows after 64px; after 64px resolution, you obtain ~85% of the range of accuracy available.

For visualization purposes, we have shown the fixations from all 8 viewers per image. The ROC performance is calculated as the average of 64 instances of 7 viewers' fixations predicting the remaining 1 viewer's fixations.

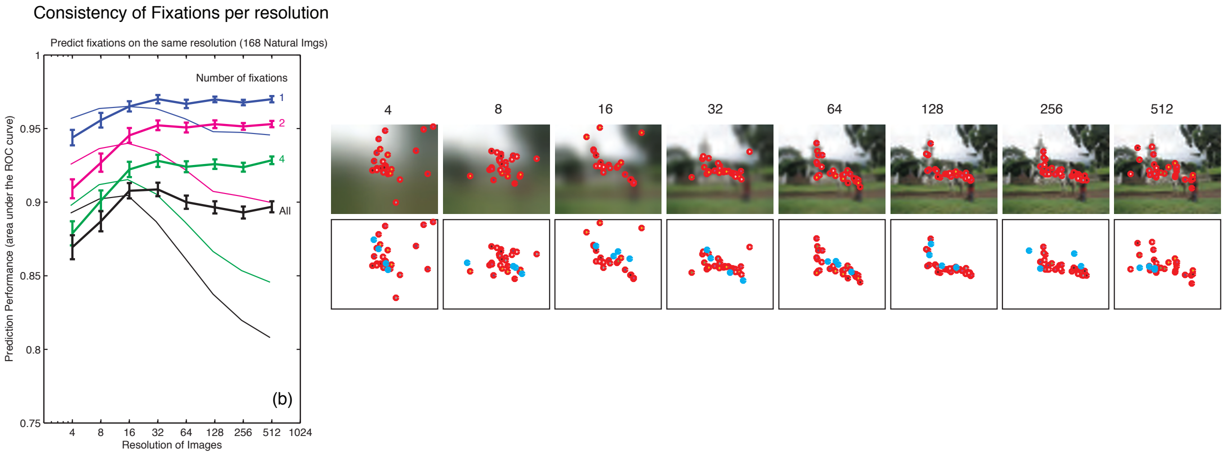


Figure 10(b) Performance on natural images. The graph on the left shows how consistent fixations on images of each resolution are; it shows how well fixations from a given image predict fixations of other viewers on the same image. This is compared to how well the center map (thin line) predicts fixations at each resolution. After 16px of resolution, human fixations outperform the center map. Human performance increases from 4-32px and after that it either plateaus (when considering the first fixations) or declines (when considering all fixations). We believe the plateau happens around the time the image is understood, and the decline happens because viewers look at the abundant extra detail of the high resolution images in their later fixations.

For visualization purposes, we have shown the fixations of 7 viewers' fixations predicting the remaining viewer's fixation. The overall ROC performance per resolution for this image is calculated as the average of the 8 of these possible instances.

fixations on 512px images (AUC=0.87) [t(63)=13.4, p<0.001]. However, fixations on lower-resolution images can still do very well at predicting fixations on the 512px image: fixations on images as low as 64px resolution predict the fixations on 512px high-resolution images with an AUC=0.90 (which is 85% of the range of performances AUC=[0.79 (4px)-0.92 (512px)]). The rate of increase in performance is highest between 4px and 64px after which the rate of increase slows down. The average prediction performance of fixations on images of 256px is equal to that of fixations on the 512px image itself.

Figure 10(a) also shows that first fixations are more consistent than all fixations. This may be because people tend to look at the most salient locations or objects in an image first (Koch & Ullman, 1985; Itti & Koch, 2000; Einhauser et al., 2008), and because they tend to look at the center first (Itti & Koch 2000, Tatler 2007, Po-He Tseng, 2009, Tatler & Vincent 2009). Earlier fixations are easier to predict in absolute terms, but they are not easier to predict relative to a relevant baseline - the center. There is a larger improvement between the baseline and the human performance for the later fixations.

Our data also shows that fixations on images above 16px predict fixations on 512px images (considering 4, AUC=0.85) significantly better than the center map (AUC=0.84) [t(63)=4.13, p<0.001]. Fixations on lower-resolution images (considering 4 fixations, 8px AUC=0.83) perform significantly worse than the center map (AUC=0.84) [t(63)=7.3, p<0.001]. The fact that humans underperform the center map is due to the number of subjects we consider and the size of the Gaussian map used to create fixation maps. We explore this further in the section on the center map.

Fixations on all resolutions predicted fixations on 512px resolutions better than chance. In the worst case, fixations on 4px image predicted fixations on 512px resolution at AUC=0.79 was significantly better than chance at AUC=0.5.

Consistency of fixations varies with resolution

Figure 10(b) shows how well fixations on a given image resolution predict fixations of other observers on the same resolution, i.e. how consistent the fixations are on each resolution. Four one-way repeated measure ANOVAs reveal a main effect of the resolution [F(7, 441)={13.8, 35.3, 29.8, 15.4} for {1, 2, 4, all} fixations with every p<0.001]. This result indicates that changes in resolution do affect fixation consistency independent of the number of fixations considered.

We investigate more fully how fixation consistency changes with increased resolution. From the graph we see that the average consistency of fixations between viewers increases with resolution from 4-32px (from AUC=0.87 to 0.92 for the 4 fixation line [t(63)=7.5, p<0.001], from AUC=0.86 to 0.90 for the all-fixation line [t(63)=6.5, p<0.001)]. Then, for the first 4 fixations, consistency between viewers plateaus after 32px, meaning observers do not get any more consistent with each other with increased resolution. This means that even through the resolution has decreased by a factor of 16 (from 512px to 32px), viewers are as consistent among each other when looking at the low-resolution images as when they look at the high-resolution images. Viewers do not need full resolution images to have consistent fixations. However, when considering all fixations, consistency seems to decrease slightly after 32px from AUC=0.90 to AUC=0.89 at 512px [t(63)=3.2, p<0.05].

While the center map outperforms humans' consistency for very low image resolutions, humans outperform the center map after 16px. At 32px, observers' fixations are significantly more consistent with each other (for the 4-fixation line AUC=0.92) than they are with the center map (AUC=0.90) [t(63)=7.6, p<0.01)] and the tendency only grows with resolution. It is interesting to note that while the consistency of the humans plateaus around 32px resolution, it does so as fixations become more spread apart. See from Figure 6 that fixations become less centered, and see from Figure 10(b) that the center map declines in performance. Despite the fact that fixations spread out, overall consistency remains about constant meaning that the performance of the humans to predict each other increases with respect to the center map baseline. Observers look at salient image features and objects when there is enough resolution to see them, rather than relying just on oculomotor biases.

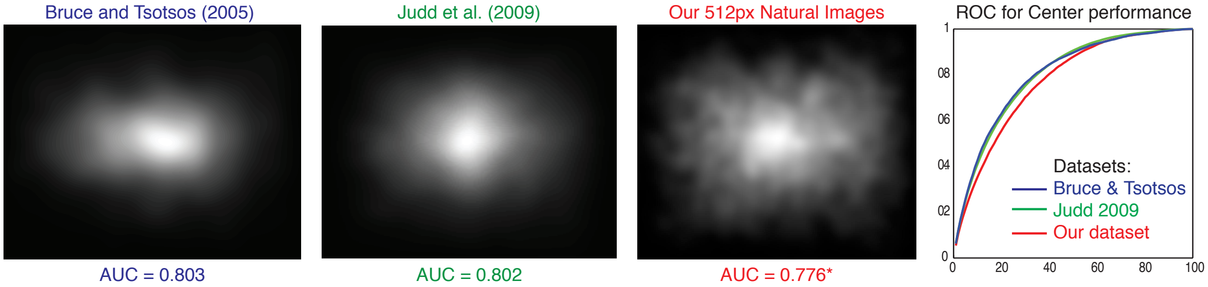


Figure 11 The average fixation maps accumulated from all fixations of all users from 3 different image databases show a clear bias towards the center of the image. Because of this strong bias, the performance of the center map to predict human fixations is quite high with an area under the ROC curve around 0.8. *The AUC for our 512px image dataset is 0.78 which differs from 0.8 (center performance of all fixations at 512px) reported in figure 10b because here we calculate center performance to predict all viewers' fixations to match calculation for other datasets and there we calculate center performance as the average of the performance for each viewer.

Performance of the center map is high because fixations are biased to the center

Figure 10 shows that performance of the center map is surprisingly high overall. For high resolution images, the center map is way above chance performance of 0.5, and as resolution decreases, the performance of the center map gets even stronger. Why is this? We examine each issue separately.

On high resolution natural images, other researchers have found that the center map produces excellent results in predicting fixations (Zhang et al., 2008; Le Meur et al., 2007) and several previous eye tracking datasets have shown that human fixations have a center bias. To compare the performance of the center map on our dataset with other datasets, we measure the performance of the center map to predict all fixations of all observers on a given image and average over all images in 3 different datasets (see Figure 11). Using this method we get an AUC of 0.78 for our 512px natural images, an AUC of 0.803 for the Bruce and Tsotsos (2005) dataset, and an AUC of 0.802 for the Judd et al. (2009) dataset. (Note that the AUC=0.78 for our dataset reported here is slightly lower than the AUC=0.8 for all fixations on 512px images reported in Figure 10(b) because there we average the performance of the center map per observer.) Ultimately, the center map performs well at predicting fixations on our database and other databases in the literature.

In Figure 12 we compare the performance of humans and the center map with two other related baselines: 1) a fixation map of a randomly selected image of the same resolution, 2) the average fixation map from all images at a given resolution. The graph shows that the average fixation map and the center map have approximately the same performance -- understandable given that the average fixation maps approximate center maps as seen in Figure 6. The average fixation map actually slightly outperforms the center map because it is slightly rectangular rather than circular and better represents fixations within a rectangular image frame. All three baselines have the same trend: they better predict fixations on low-resolution images than on higher resolution images though their absolute values are different. This indicates that fixations across images are more similar and more centered at lower resolutions than at higher resolutions.

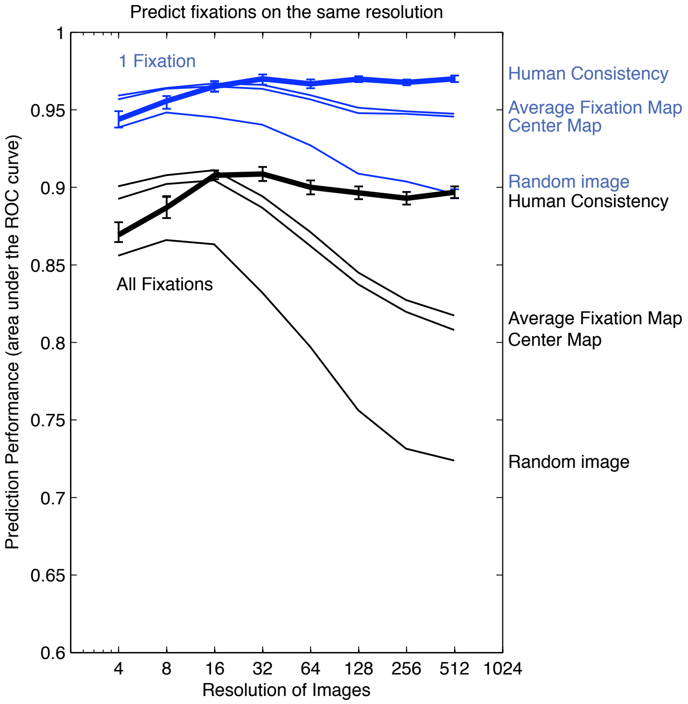


Figure 12 Human consistency compared to three baselines. The performance of all three baselines--the center map, average fixation maps and random image maps--have the same trend though their absolute values are different. Performance declines with increased resolution. This indicates that fixations across images are more similar and more centered at lower resolutions than at higher resolutions.

Researchers have studied why fixations are biased towards the center of the image and show that it is due in part to photographic bias (people place objects of interest near the center of their photo composition), viewing strategy (tendency to expect objects of interest in the center), orbital reserve (tendency to look straight ahead), and screen center (tendency to look at the center of a display) (Tatler, 2007; Tatler & Vincent 2009; Tseng et al. 2009). The photographic bias plays a large role in the center bias of fixations for high-resolution images, but must have a much weaker role on low-resolution images where the principal objects of interest are not recognizable. On low-resolution images, the effects of viewing strategy, orbital reserve and screen center account for the central fixation bias.

The performance of the center map helps us interpret human consistency performance: 1) One of the reasons human consistency is high around 16-32px is definitely influenced by the fact that fixation patterns overall are quite similar on these images. 2) Though absolute performance of human consistency remains somewhat constant from 32-512px, baselines go down and the relative performance of human consistency over any baseline improves dramatically with resolution (Figure 10b). We hypothesize that people are consistent at around 16-32px because fixations across different images look similar (near the center), whereas people are consistent at higher resolution because they look at the same (not necessarily centered) salient locations specific to the image.

One curiosity remains: why is the center map performance higher than human performance on low resolution natural images? Human performance is due to both 1) the number of subjects considered and 2) the size of the Gaussian filter we use to create the fixation map, and these components become more important at low resolution images where fixations are less due to actual salient objects and more due to human biases. The more subjects we consider, the more the human fixation map will match the average fixation or center map and the closer the performance will be to the center map performance. Secondly, the larger the Gaussian filter we consider for creating the fixation map, the more the fixation map matches the center map. When the Gaussian filter is enlarged to 1 to 2 cycles per image, the performance increases for the resolutions 4 and 8 and approaches the performance of the center map for those resolutions. This is reasonable given that the center map is a 1 cycle per image Gaussian located at the center of the image. At 16px and above, the highest performing Gaussian filter was 4 cycles per image, though it was only slightly higher than our original Gaussian of 8 cycles per image. For simplicity, we use 8 cycles per image for all resolutions.

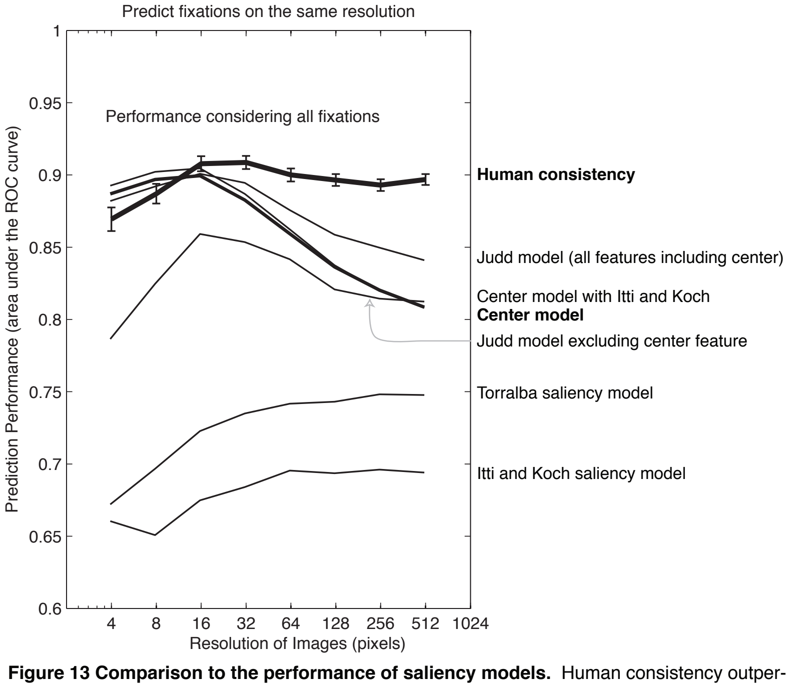
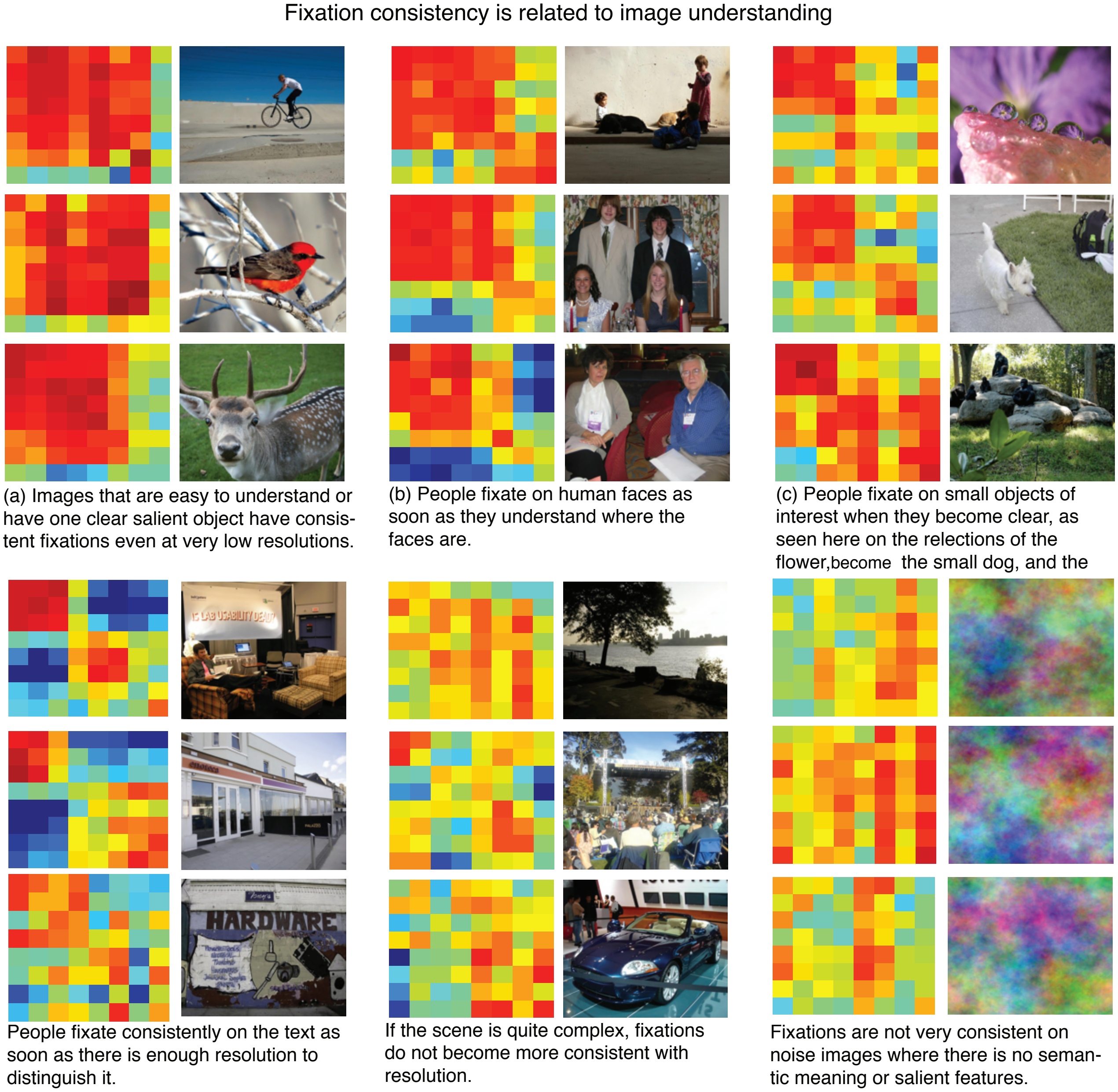


Figure 13 Comparison to the performance of saliency models. Human consistency outperforms saliency models in predicting fixations for all models. The center outperforms all models that do not include the distance to the center of the image as a feature of the model.

Human consistency and center performance outperform most saliency models

We compare the human performance and center performance with some state of the art saliency models. Figure 13 shows that for high resolution images the center model outperforms the Itti and Koch model (Itti and Koch, 2000), Torralba's model (Torralba et al. 2006) and the Judd model (Judd et al. 2009) without the center feature. What is interesting from this analysis is how the performance of the different models change as the resolution of the images increases. In general, saliency models tend to increase in performance while the center decreases in performance as resolution increases. This is not entirely the case for the Judd model without the center feature which rises and then dips with increased resolution. The models that outperform the center model for most resolutions are models that include the center map as a feature in the model. This is the case of the Judd et al. 2009 model and the model which combines the Itti and Koch model with the center map. We include the performance of these models for reference, but studying their behavior in depth is beyond the scope of this paper.


Fixation consistency is related to image understanding

(a) Images that are easy to understand or have one clear salient object have consistent fixations even at very low resolutions.

(b) People fixate on human faces as soon as they understand where the faces are.

(c) People fixate on small objects of interest when they become clear, as seen here on the relections of the flower,become the small dog, and the

People fixate consistently on the text as soon as there is enough resolution to distinguish it.

If the scene is quite complex, fixations do not become more consistent with resolution.

Fixations are not very consistent on noise images where there is no semantic meaning or salient features.

Figure 14 These images and their corresponding prediction matrices (for the first 4 fixations) show that fixation consistency is related to image understanding. The higher the consistency, the redder the pixel. As soon as the image is understood, people look at the same locations. Different images require more or less resolution to be understood; the more complex the image, the more resolution is needed.

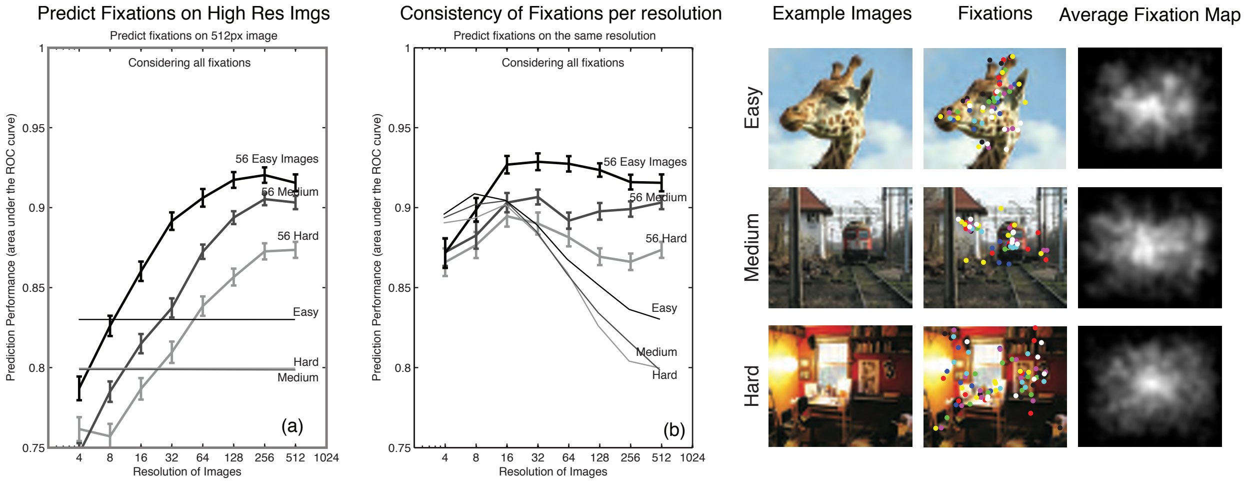


Figure 15 Performance on easy, medium and hard images. Here we see the all-fixation trend line separated for different types of images: easy, medium and difficult images to understand. When images are "easy", consistency between different resolutions and consistency on a given resolution are much higher than for "hard" images. In addition, consistency peaks and then declines more strongly for the hard images as resolution increases. In this case, fixations are consistent at low resolution and then become less consistent as the small, complex details of the hard images become visible.

Image complexity affects fixation consistency

What causes the plateau or decrease in fixation consistency to appear after 16-64px of resolution? We discuss two major affects.

Firstly, the range of 16-64px may be the threshold at which images are understood (Torralba, 2009). With less resolution, an image has no semantic meaning and people look randomly at bottom-up salient locations. After around 16-64px, there is enough low spatial frequency in the image that viewers can understand the gist of the scene (Bar, 2004, 2007). This global information informs about scene category (McCotter et al.,2005) and informs viewers of which objects are likely to be in the scene and where (Loftus & Machworth, 1978; Biederman et al., 1982; De Graef et al., 1990; Henderson et al., 1999). It also may be that this is where the primary objects of the scene become visible and thus attract fixations directly (Einhäuser et al., 2008; Elazary & Itti, 2008). One can see this effect in Figure 14 where the prediction matrices for specific images have a "square-pattern" indicating a large jump and then plateau in performance after a given resolution. The resolution at which this jump occurs depends on the content and complexity of the image. Some images are understood early as in 14(a) and others later as in 14(c) and (d).

Secondly, when considering all fixations, images at 16-64px have a peak in fixation consistency because it is a threshold at which the image is understandable but not yet full of tiny image details. As image resolution increases, more small details become visible and attract fixations. This is particularly true for images that are more complex in nature. As complexity increases, more resolution is needed to understand what the image contains, and at high resolution there are more things to look at.

Figure 15 shows the fixation consistency of all fixations of the three subsets of natural images. A 3x8 (image complexity type x resolution) repeated measures ANOVA on Figure 15(b) revealed main effects of the image complexity type [F(2, 126) = 202.9, p <0.001], and resolution [F(7, 441)= 15.4, p<0.001], and a significant interaction [F(14, 1512)= 5.6, p<0.001]. This result indicates that increased resolution affects fixation consistency, and that the type of image also affects the consistency with easier images having higher consistency. Interestingly, there is also an interaction between the two components showing that predictability goes down as resolution goes up, but this happens faster as the image type gets harder. The decline in fixation consistency appears most strongly for the hard images. These are the images where extra resolution brings extra details which spread fixations out and lead to less consistency. This effect is less evident on the medium and easy complexity images. On easy images, fixations go directly to the main object of interest at both low and high resolution because there is often no other details to look at (consider the image of the head of the giraffe or the girl's face in Figures 2 and 3).

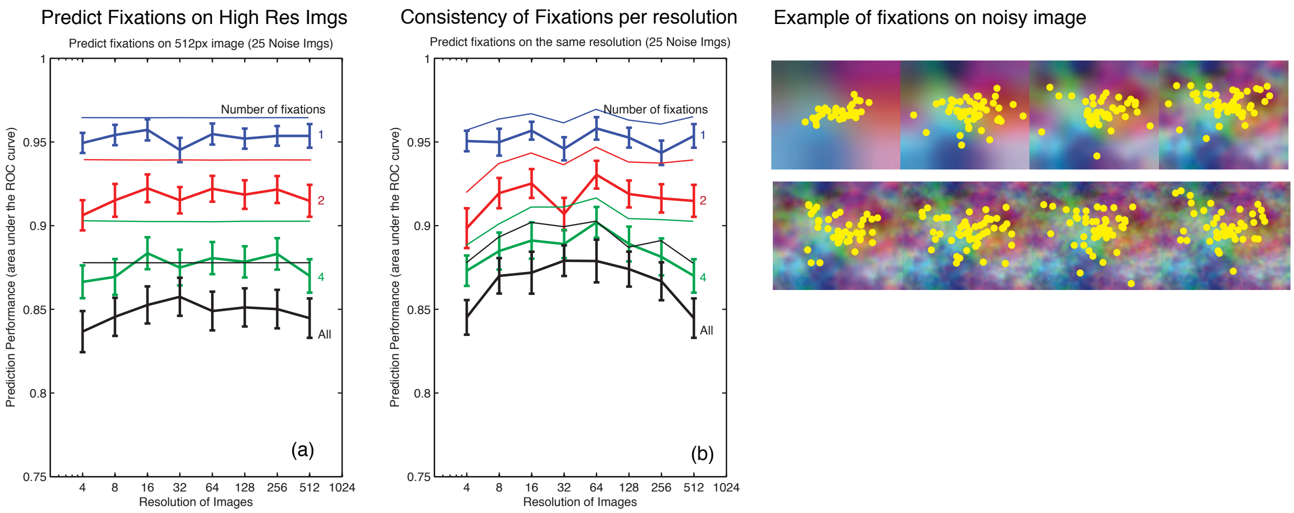


Figure 16 Performance on noise images. As a control, we plot the performance of fixations on noise images. Compare these graphs to the performance of natural images. In graph (a) we find that as resolution increases, human fixations do not get more accurate at predicting fixations on high resolution images. In addition, humans never out perform the gaussian center. From graph (b) we see that no particular resolution has more consistent fixations than the others, and once again the gaussian center better predicts fixations than humans for all resolutions.

Consistency of fixations on noise images is poor

For comparison, Figure 16 (a) (b) show fixation consistency for the pink noise images. We see that for noise images the center map is the best performer: it out-predicts observers on all numbers of fixations. This would change as we add more viewers to the experiment; more viewers would mean that human fixation maps would be closer to the average fixation map, and thus more closely match the predictions of the center map.

Additionally, we see that the curves are mostly flat. This means that predicting fixations on high-resolution images does not increase with fixations from images of increasing resolution; fixations from a low-resolution image are just as good as fixations from a high-resolution image at predicting fixations on high-resolution images.

These noise images, despite matching natural images statistics, have lost natural image semantics and salient features. When these are removed, fixations seem only to be consistent with the center map.

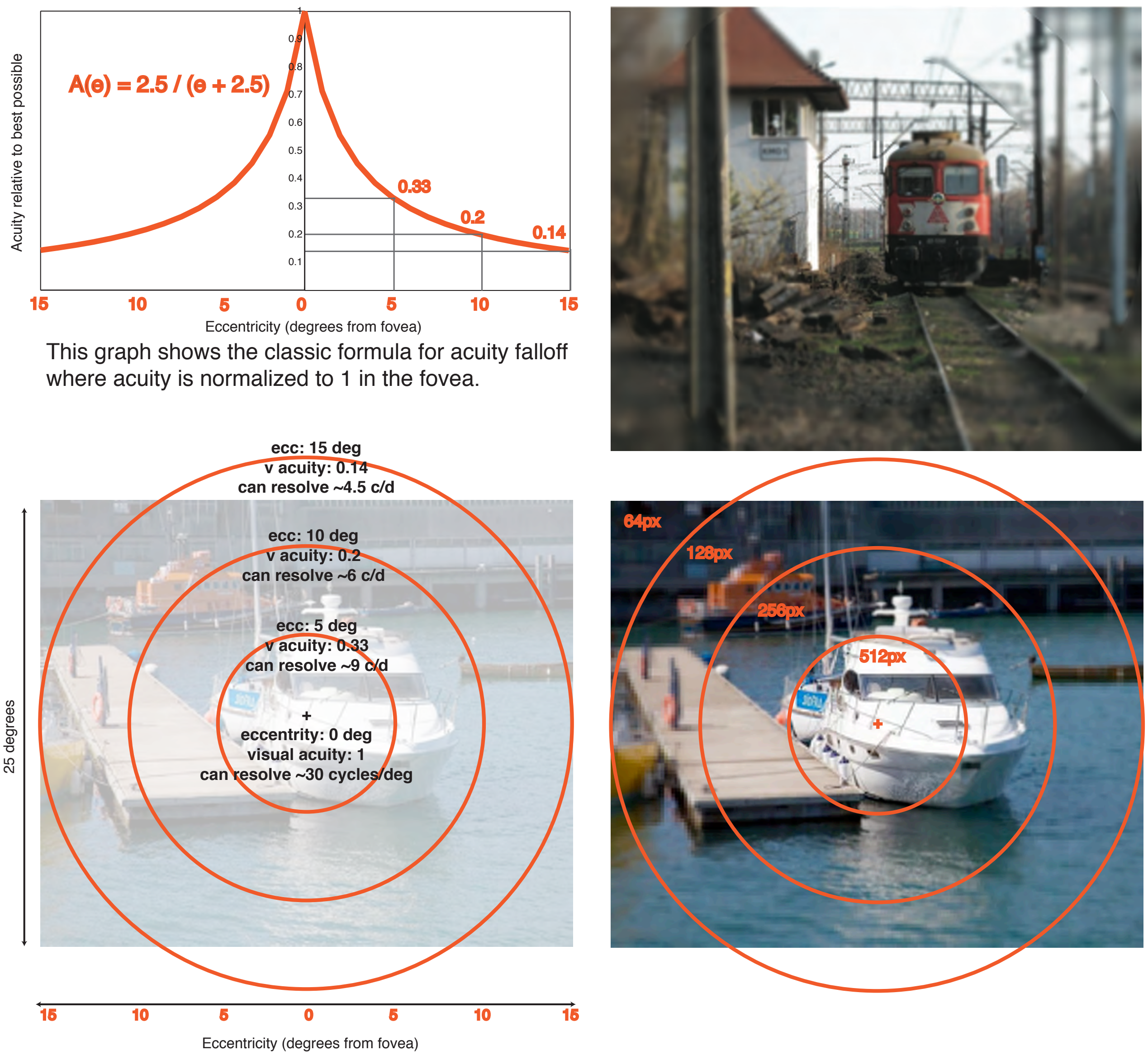
Falloff in visual acuity may affect fixations

It is well understood that the spatial resolution of the human visual system decreases dramatically away from the point of gaze (Anstis, 1998). The resolution cutoff is reduced at a factor of 2 at 2.5 degrees from the point of fixation, and by a factor of 10 at 20 degrees as seen in Figure 17 (Geisler and Perry, 1998). Since our images extend an angle of 30 degrees, we can approximate the relative visual acuity, and therefore the number of cycles per degree that a viewer can resolve, at different locations in the image.

Given a viewer is fixating at the center of the image, he can resolve 30 cycles per degree (assuming he has normal 20/20 vision) at the center. At 5 degrees of eccentricity, the viewer has 33% of their original visual acuity and can resolve 9 cycles per degree, corresponding to a 256px image. At 15 degrees of visual angle, or at the horizontal edges of the image, the viewer has 14% of visual acuity and can resolve about 4 cycles per degree, corresponding to 128 px.

When a viewer looks at a 512px image, he cannot resolve all the available details in the periphery. When he looks at an image of 64px or below, he is able to resolve all image information even in the periphery because the entire image is below 2 cycles per degree. At this resolution, there is no difference between the center and the periphery in what a viewer can resolve.

We hypothesize that this could have the following implications: 1) For high resolution images, viewers must move their eyes (actively saccade and fixate) in order to resolve different parts of the image. This might not be the case for resolutions 64px and below. 2) The center performs very well for images of 32px and below. In these cases, the resolution is so low (everything is less than 1 cycle per degree) that the entire image can be resolved using peripheral vision; viewers don't need to move their eyes to distinguish details.



This graph shows the classic formula for acuity falloff where acuity is normalized to 1 in the fovea.

Figure 17 Modeling visual acuity falloff The images on the right simulate the acuity falloff of the human visual system given a fixation at the center of the image. They have high resolution (512px or 16 cycles per deg) at the center and low resolution (64px or 2 cycles per deg) on the edge. They should be seen so that they extend 30 degrees of visual angle. If they are printed on a paper at 3in wide, they should be viewed about 6in away.

This could account for the very high performance of the center model on low resolution images. As the cycles per degree get higher, eyes move away from the center in order to resolve the image details.

Conclusions

We draw the following conclusions from this work:

Firstly, fixations from a specific resolution image are best predicted by fixations on the same resolution image. However, fixations from lower-resolution images can also quite reliably predict fixations on higher-resolution images. Prediction performance drops slowly with decreasing resolution until 64px after which it drops more quickly.

Secondly, human fixations become more consistent as resolution increases until around 16-64px after which consistency remains relatively constant despite the spread of fixations away from the center. We hypothesize that consistency stays strong despite the decreasing center bias of fixations because the image becomes understandable and viewers start to look consistently at the saliency objects or locations.

Thirdly, there is a significant bias for the fixations on all images to be towards the center. For high resolution images, the area under the ROC curve for the center map predicting all fixations is ~0.8 and agrees with the performance of other datasets in the literature. As resolution decreases, fixations get more concentrated at the center and the performance of the center map increases. This trend agrees with two other baselines: the performance of randomly shuffled image maps and average fixation maps and indicates that fixations across images are more similar at lower resolutions than at higher resolutions.

Fourth, humans predicting other human fixations outperform any model of saliency that aim to predict where people look. In addition, the center map also outperforms any model that does not include the center as a feature of the model.

Finally, fixation consistency is directly related to the complexity of images. Images which are easier to understand have higher consistency overall and remain high with increasing resolution. Images that are more complex and require more resolution to be understood often have lower overall fixation consistency and decrease in fixation consistency with resolution. We hypothesize that this is because later fixations get diverted to small details.

These trends provide insight into how the human brain allocates attention when regarding images. This experiment shows that viewers are consistent about where they look on low-resolution images and are also looking at locations consistent with where they look on high-resolution images. These findings suggest that working with fixations on mid-resolution images instead of on high-resolution images could be both perceptually adequate at the same time as being computationally attractive.

This result would be useful for real-time image analysis applications, such as robotic vision, which uses a saliency map to prioritize image locations for further processing. Instead of computing saliency on full-resolution images, the preprocess could be sped up significantly by working instead with low-resolution images.

For future work, it would be interesting to better understand how different computational saliency models predict fixations on varying-resolution images differently and how this depends on whether the saliency model is a bottom-up model or includes top-down image cues. In addition it would be interesting to run an experiment where subjects are both explicitly asked about their understanding of an image and tracked for eye-movements. This could lead to results which more directly support the hypothesis that eye movements are under heavy influence of image understanding.

Acknowledgments

This research was supported by NSF CAREER awards 0447561 and IIS 0747120. Frédo Durand acknowledges a Microsoft Research New Faculty Fellowship, a Sloan Fellowship, Royal Dutch Shell, the Quanta T-Party, and the MIT-Singapore GAMBIT lab. We thank Aude Oliva, Barbara Hidalgo-Sotelo and Krista Ehinger for their help and comments and Aude Oliva for the use of her eye tracker.

Commercial relationships: none.
Corresponding author: Tilke Judd.
Email: tjudd@csail.mit.edu.
Address: 32 Vassar Street, Cambridge, MA 02139.

References

Anstis, Stuart. (1998). Picturing peripheral acuity. *Perception*, 27, 817-825.

Avraham and M. Lindenbaum. (2009) Esaliency: Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1).

Bachmann, T. (1991). Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology, 3*, 85–103.

Bar, M. (2004). Visual objects in context. *Nature Neuroscience Reviews 5*, 617–629.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences 11*, 280–289.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 143-177.

Bruce, N. & Tsotsos, J.K. (2009) Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9(3)*:1–24, 3.

Bruce, N. & Tsotsos, J. K. (2006). Saliency Based on Information Maximization. *Advances in Neural Information Processing Systems, 18*, 155-162.

Buswell G. T. (1935). *How people look at pictures. A study of the psychology of perception in art.* Chicago: University of Chicago Press.

Castelhano, M.S. & Henderson, J.M. (2008). The influence of color on perception of scene gist. Journal of Experimental Psychology: *Human Perception and Performance 34*, 660–675.

Cerf M., Harel J., Einhäuser W., Koch C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems, 20*, 241–248.

De Graef, P., Christiaens, D., & d'Ydevalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research, 52*, 317-329.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 Scenes: A combined source model of eye guidance. *Visual Cognition, 17*, 945-978.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8(14),* 18: 11-26.

Einhäuser W., Rutishauser U., Koch C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8,* (2):2

Elazary, L., and Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8(3):3*, 1-15.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory of gist. *Journal of Experimental Psychology: General 108,* 316–355.

Geisler, W.S. and Perry, J.S. (1998) A real-time foveated multi-resolution system for low-bandwidth video communication. In B. Rogowitz and T. Pappas (Eds.), *Human Vision and Electronic Imaging, SPIE Proceedings*, 3299, 294-305.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in Neural Information Processing Systems, 19,* 545-552.

Harmon, L.D. & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional filtered noise. *Science, 180,* 1194–1197.

Henderson J. M., Brockmole J. R., Castelhano M. S., Mack M. van R., Fischer, M., Murray, W., Hill R. (2006). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movement research: Insights into mind and brain.* (pp. 537–562). Oxford: Elsevier.

Henderson, J. M., Weeks, P. A. Jr., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 210-228.

Henderson J. M., McClure K. K., Pierce S., Schrock G. (1997). Object identification without foveal vision: Evidence from an artificial scotoma paradigm. Perception & Psychophysics, 59, 323–346.

Hershler O., Hochstein S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research, 45,* 1707–1724.

Hershler O., Hochstein S. (2006). With a careful look: Still no low-level confound to face pop-out. *Vision Research, 46,* 3028–3035.

Hou, X, Zhang, L. (2007). Saliency detection: A spectral residual approach. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 0:1-8.

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489-1506.

Itti, L., Koch, C., & Niebur, E. (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision, Vol. 20(11),* pp. 1254.

Judd, T., Durand, F, & Torralba, A. (2009). Learning to Predict Where Humans Look. *IEEE Conference on Computer Vision (ICCV).*

Kayser C., Nielsen K. J., Logothetis N. K. (2006). Fixations in natural scenes: Interaction of image structure and image content. *Vision Research, 46,* 2535–2545.

Kienzle, W., Wichmann, F. A., Scholkopf B., and Franz, M. O. (2007) A nonparametric approach to bottom-up visual saliency. In B. Scholkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 689–696. MIT Press.

Koch and Ullman (1985) Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219-227.

Le Meur, Olivier, Le Callet, Patrick, Barba, Dominique, Predicting visual fixations on video based on low-level visual features. Vision Research 2007 Volume 47 Issue 19: 2483-2498.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6(1),* 9-16.

Loftus, G. R. & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 565-572.

Mannan, Ruddock & Wooding (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9(3):* 363-386.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2005). The use of visual information in natural scenes. *Visual Cognition, 12,* 938-953.

Navalpakkam V., Itti L. (2007). Search goal tunes visual features optimally. *Neuron, 53,* 605–617.

Neider, M. B., and Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research, 46,* 614-621.

Niebur and Koch, (1995). *The Attentive Brain*, chapter Computational architectures for attention,163-186. Cambridge MA: MIT Press.

Oliva, A. (2005). Gist of the scene. *In The Encyclopedia of Neurobiology of Attention,* ed. Itti, L., Rees, G. & Tsotsos, J.K., pp. 251–256. San Diego, CA: Elsevier.

Oliva, A. & Schyns, P.G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology, 41,* 176–210.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42(3),* 145–175.

Oliva, A., Torralba, A., Castelhano, M., & Henderson J. (2003). Top-down control of visual attention in object detection. *Proceedings of the 2003 International Conference on Image Processing.* (p.1).

Parkhurst, D. J. & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16(2),* 125-154.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107-123.

Peters R. J., Iyer A., Itti L., Koch C. (2005). Components of bottom–up gaze allocation in natural images. *Vision Research, 45,* 2397–2416.

Privitera C. M., Stark L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 970–982.

Potter, M.C. (1975). Meaning in visual scenes. *Science 187*, 965–966.

Potter, M. C., and Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81,* 10-15.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., and Chua, T.-S. (2010). An eye fixation database for saliency detection in images. *European Conference on Computer Vision* (ECCV).

Rao R. P., Zelinsky G. J., Hayhoe M. M., Ballard D. H. (2002). Eye movements in iconic visual search. *Vision Research, 42,* 1447–1463.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 1-17.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research 39,* 19:3157-3163.

Schyns, P.G. & Oliva, A. (1997). Flexible, diagnostically-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception, 26,* 1027–1038.

Simoncelli, Eero, *The Steerable Pyramid Toolbox*, Available at http://www.cns.nyu.edu/~eero/steerpyr/.

Sinha, P., Balas, B.J., Ostrovsky, Y. & Russell, R. (2006). Face recognition by humans: 19 results all computer vision researchers should know about. *Proceedings of the IEEE, 94 (No. 11),* 1948–1962.

Tatler B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, (14):4, 1–17.

Tatler, B. W., Baddeley, R. J., Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5), 643-659.

Tatler, B. W. & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17(6-7),* 1029-1054.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of Optical Society of America A. Special Issue on Bayesian and Statistical Approaches to Vision, 20(7),* 1407-1418

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience*, *26*, 123-131.

Torralba A., Oliva A., Castelhano M. S., Henderson J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113,* 766–786.

Tseng, Carmi, Cameron, Munoz, Itti, (2009) Quantifying center bias of observers in free viewing of dynamic natural scenes, *Journal of Vision, 9,* (7):4, 1–16.

Tsotsos J. K., Culhane S., Wai W., Lai Y., Davis N., Nuflo F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78,* 507–547.

Underwood G., Foulsham T., van Loon E., Humphreys L., Bloyce J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology, 18,* 321–343.

VanRullen R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Research, 46,* 3017–3027.

Velichkovsky, B., Pomplun, M., & Rieser, J. (1996). Attention and communication: Eye-movement-based research paradigms. In W. H. Zangemeister, H. S. Stiehl, & C. Freksa (Eds.), *Visual attention & cognition (pp. 125-154).* Amsterdam: Elsevier.

van Zoest, W., Donk, M., & Theeuwes, J. (2004). The role of stimulus-driven and goal-driven control in saccadic visual selection. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 746–759.

William J. (1981) Principles of Psychology. Cambridge, MA: Harvard University Press. Originally published in 1890.

Wolfe, J.M. (1998). Visual memory: What do you know about what you saw? *Current Biology 8,* R303–R304.

Yarbus A. L. (1967). *Eye movements and vision.* New York: Plenum.

Zhang L., Tong, M. H., Marks, T. K., Shan H., and Cottrell, G. W. (2008) SUN: A Bayesian framework for saliency using natural statistics. *J. Vis., 8(7):*1–20, 12.